FIGURE 1

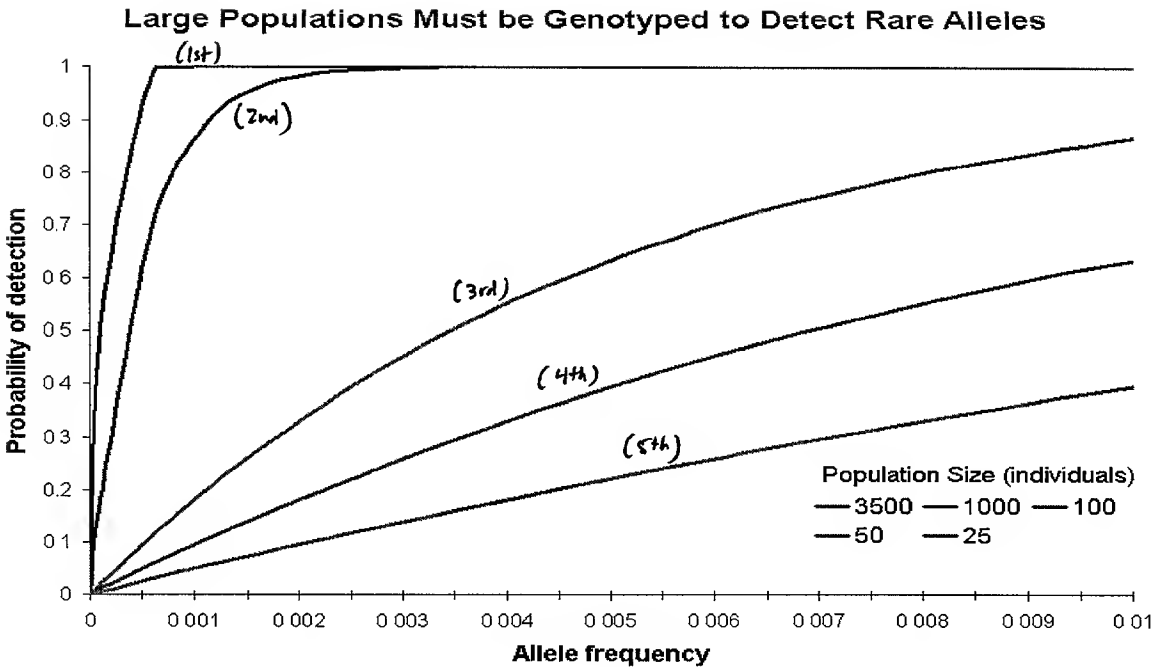
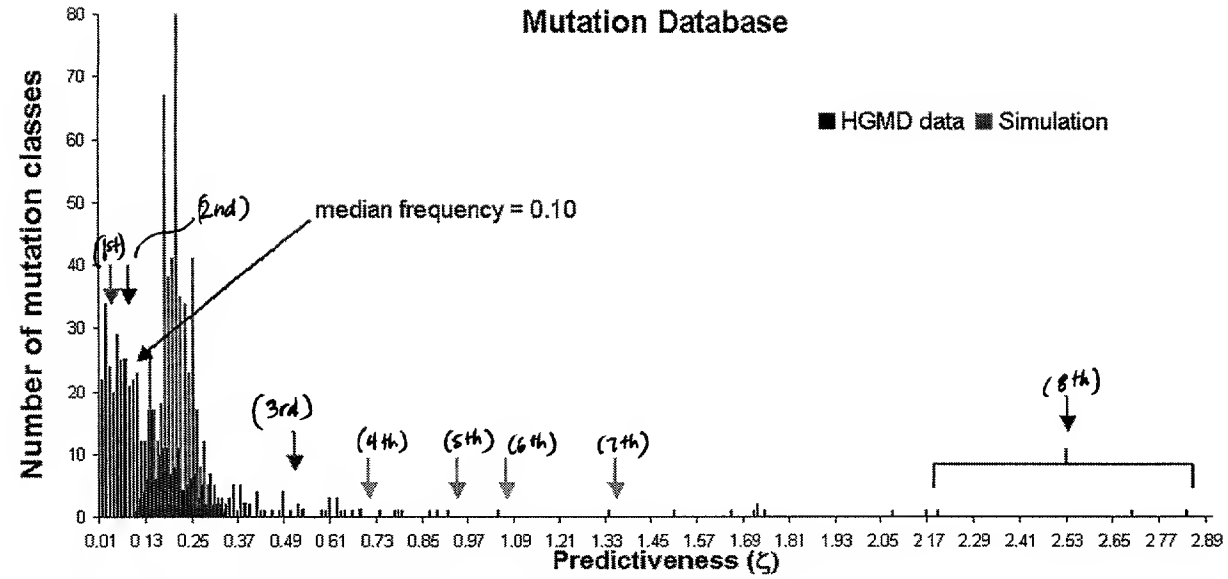
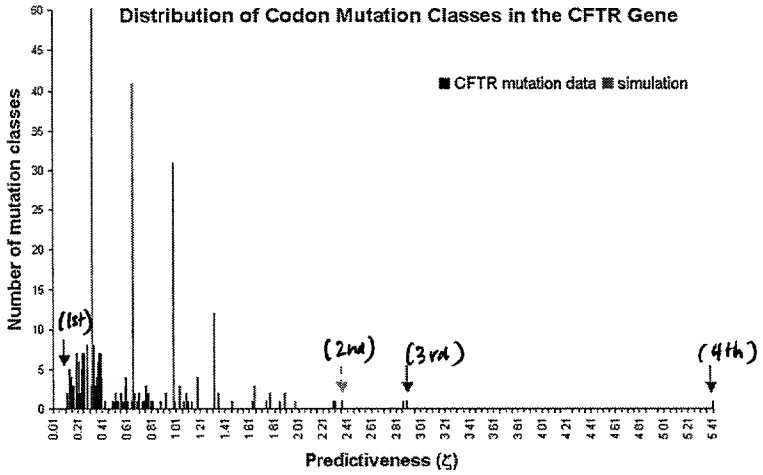

Large Populations Must be Genotyped to Detect Rare Alleles
Probability of detection
Allele frequency
(1st)
(2nd)
(3rd)
(4th)
(5th)
Population Size (individuals)
3500
1000
100
50
25
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
0.01

FIGURE 2


Mutation Database
HGMD data
Simulation
Number of mutation classes
median frequency = 0.10
(1st)
(2nd)
(3rd)
(4th)
(5th)
(6th)
(7th)
(8th)
Predictiveness (ζ)


Distribution of Codon Mutation Classes in the CFTR Gene
CFTR mutation data
simulation
Number of mutation classes
(1st)
(2nd)
(3rd)
(4th)
Predictiveness (ζ)

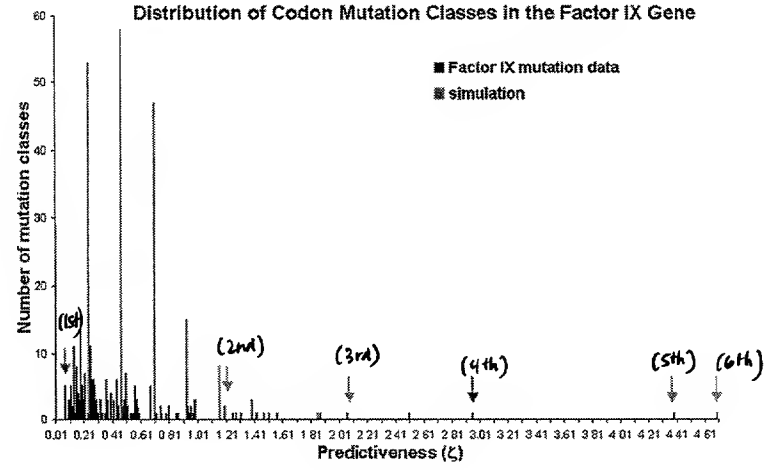

Distribution of Codon Mutation Classes in the Factor IX Gene
Factor IX mutation data
simulation
Number of mutation classes
(1st)
(2nd)
(3rd)
(4th)
(5th)
(6th)
Predictiveness (ζ)

FIGURE 3

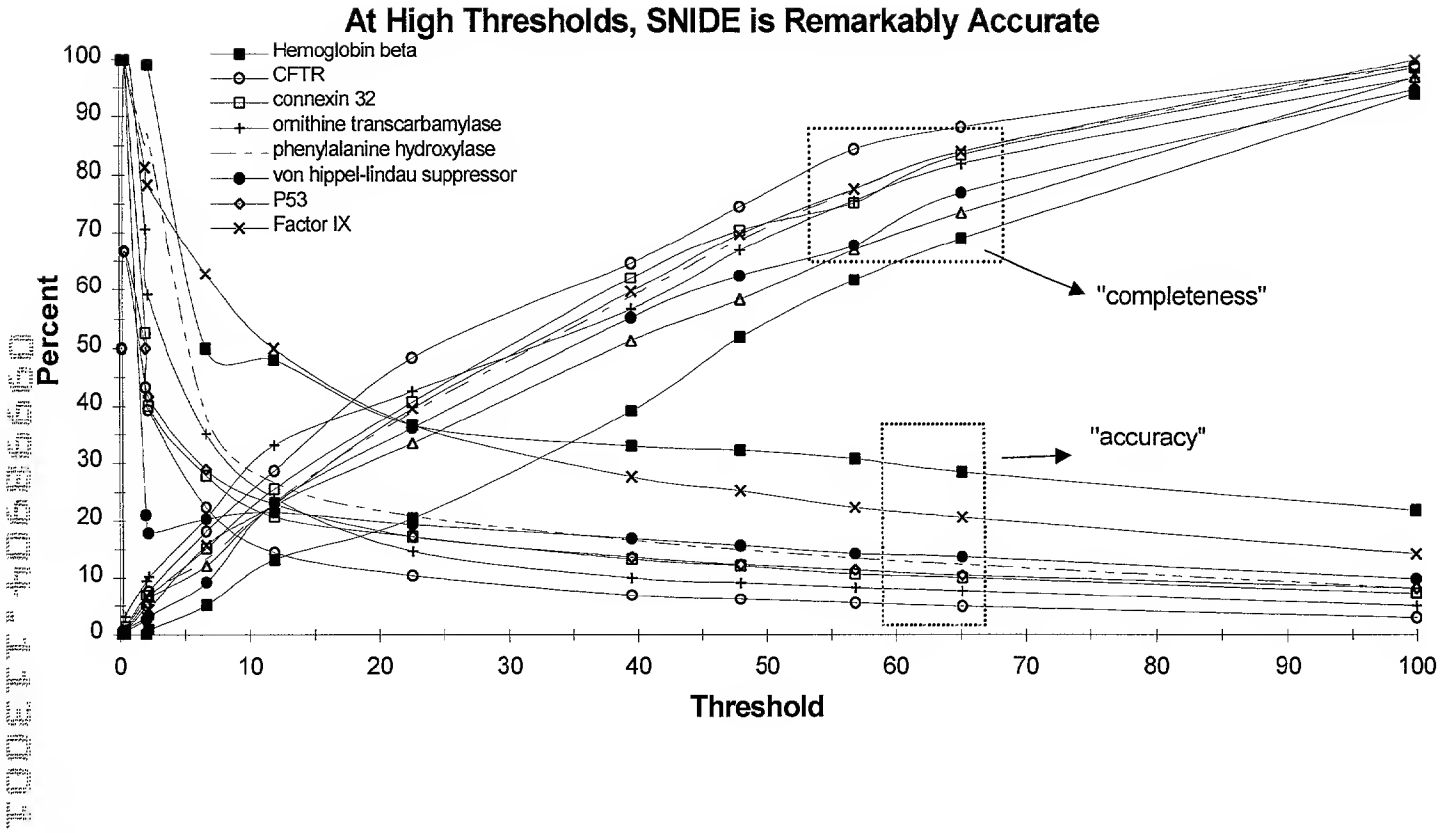

At High Thresholds, SNIDE is Remarkably Accurate
Hemoglobin beta
CFTR
connexin 32
ornithine transcarbamylase
phenylalanine hydroxylase
von hippel-lindau suppressor
P53
Factor IX
"completeness"
"accuracy"
Percent
Threshold
0
10
20
30
40
50
60
70
80
90
100

FIGURE 4

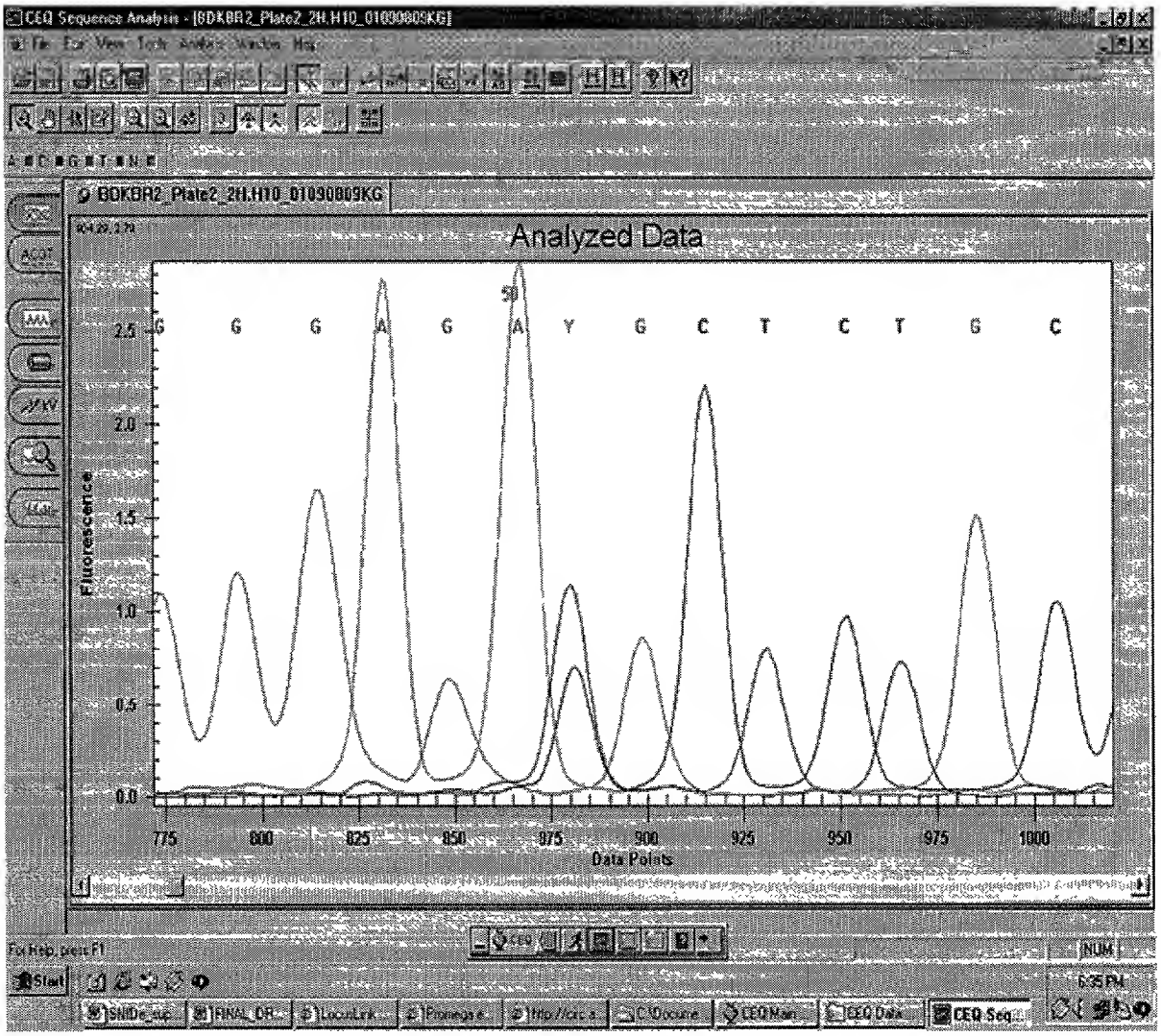
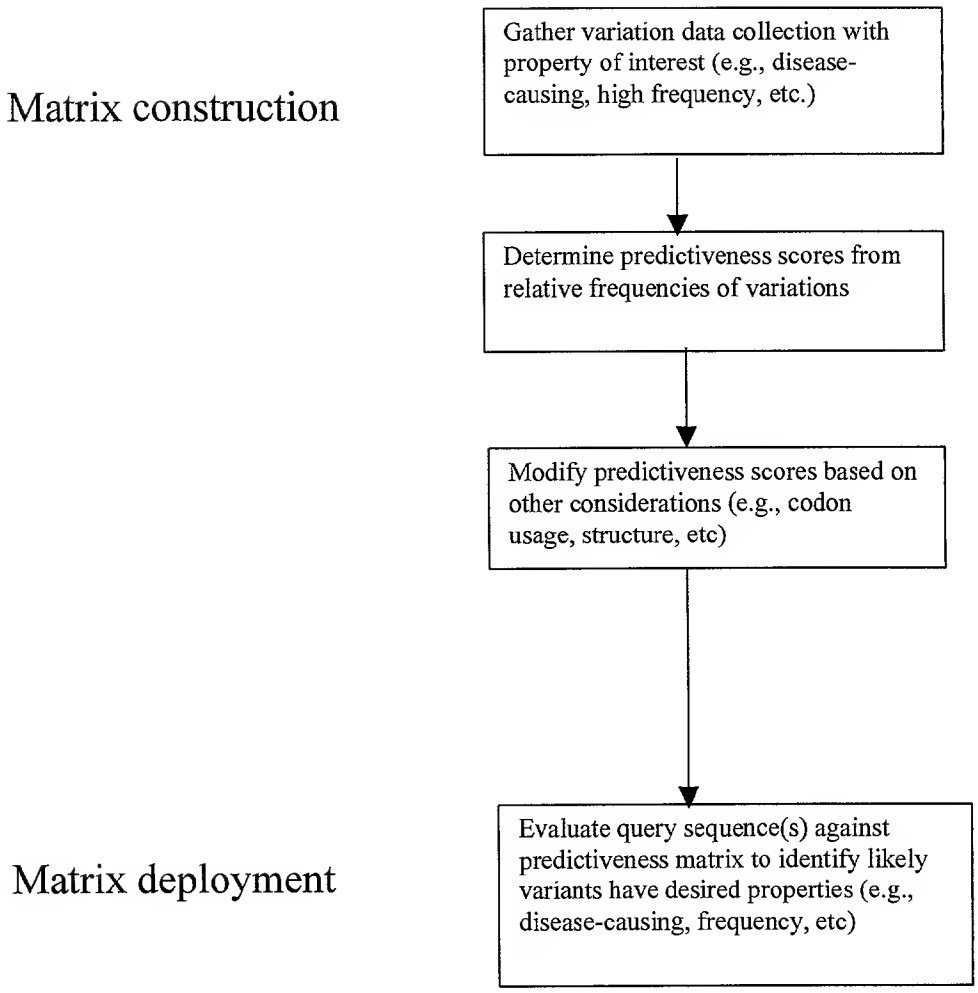

CEQ Sequence Analysis - [BDKBR2_Plate2_2H.H10_01030809KG]
BDKBR2_Plate2_2H.H10_01030809KG
Analyzed Data
G G G A G A Y G C T C T G C
Fluorescence
2.5
2.0
1.5
1.0
0.5
0.0
775
800
825
850
875
900
925
950
975
1000
Data Points
NUM
6:35 PM

FIGURE 5


Matrix construction
Gather variation data collection with property of interest (e.g., disease-causing, high frequency, etc.)
Determine predictiveness scores from relative frequencies of variations
Modify predictiveness scores based on other considerations (e.g., codon usage, structure, etc)
Matrix deployment
Evaluate query sequence(s) against predictiveness matrix to identify likely variants have desired properties (e.g., disease-causing, frequency, etc)